2014

Bayesian Model-building in Phylogenetics

Bradley Nelson
Louisiana State University and Agricultural and Mechanical College

BAYESIAN MODEL-BUILDING IN PHYLOGENETICS

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Biological Sciences

by
Bradley J. Nelson
B.S., Seattle Pacific University, 2010
August 2014

Acknowledgments

This thesis was made possible through the advice and knowledge of Dr. Jeremy M. Brown, helpful comments from committee members Dr. Bret Elderd and Dr. Mike Hellberg, the patient support of my partner, Katherine M. M. Steensma, and my parents, whose love and generosity helped make me the person I am today.

Table of Contents

Abstract

DNA sequencing costs have decreased dramatically over recent decades, resulting in a flood of phylogenetic information available to researchers. While it is often assumed that additional data will lead to more accurate conclusions, it also raises a number of problems for phylogeneticists, including mundane computational issues such as data management and complex statistical problems such as obtaining a single species tree from multiple conflicting gene trees. Developing new methods to make better use of existing data and probe the causes of conflicting signal will be necessary to confidently resolve phylogenies in the genomic era.

Here, we examine two current problems in statistical phylogenetics and attempt to address them in a Bayesian framework. The first problem involves inflated tree lengths in Bayesian phylogenies, which can be an order of magnitude longer than maximum likelihood estimates. We developed EmpPrior, a program which queries TreeBASE for datasets similar to the focal data, then estimates parameters from each dataset to inform priors on the focal data. This approach greatly improves the tree length credible intervals in four exemplar datasets and, when combined with other approaches such as the use of a compound Dirichlet prior on tree length, can nearly eliminate the problem of inflated trees.

The second problem involves incongruence between morphological and molecular phylogenies in squamates. Here, we use posterior prediction with inferential test statistics to investigate whether systematic error may be biasing inference in the molecular data. While we detected some model violation in most of the 44 genes, the genes with the most model violation were more distant from the molecular phylogeny. This suggests that model violation is not a major source of error

in the molecular data. Hence, the source of incongruence between the molecular and morphological squamate topologies remains unknown.

In both problems, we found that incorporating tools such as informed priors and posterior prediction from Bayesian statistical literature into phylogenetic analyses can improve results and help uncover why different datasets lead to conflicting topologies. As phylogenetic datasets continue to grow, using methodological best practices will only become more important if we want to have confidence in our conclusions.

Chapter 1
Introduction

Understanding the evolutionary relationships between organisms is one of the fundamental goals of biology. The field of phylogenetics, which involves the construction of evolutionary trees from molecular data, has resolved many nodes in the tree of life with confidence, but the relationships between many clades remains uncertain. Recent developments in sequencing technology have made it financially feasible to determine some of these relationships, but with the massive influx of new data come a host of computational and statistical challenges.

The rise of next-generation sequencing over the past 15 years has transformed genetic research. The cost of sequencing a human genome has fallen from nearly $100 million in 2001 to less than $5000 in 2014 (Wetterstrand). This change has opened up a plethora of previously infeasible research opportunities, including genome-wide association studies, transcriptomics, and resolving deep divergences in phylogenetics. Such studies are beginning to unravel the genetic basis of complex diseases, how gene expression varies with environment, and the structure of the tree of life. But along with these opportunities come a host of new challenges researchers must grapple with.

Analyzing next-generation sequence information often requires dealing with big data. Sequencing cost has decreased at a faster rate than Moores law over the past 15 years, which means that the cost of analyzing genetic datasets has increased relative to the cost of generating those datasets. The full 1000 Genomes dataset is approximately 200 TB, putting it well beyond the range of what can be contained on a single desktop computer. Dealing with datasets of this size involves a set of

challenges which may be unfamiliar to traditional biologists, including organizing large analyses on a remote computer cluster, implementing software that makes efficient use of disk space and computation time, and summarizing the results of these analyses in a way that is biologically meaningful. But challenges with next-generation sequencing are not solely limited to issues with big data.

Another set of issues is statistical in nature. For phylogenetics in particular, traditional datasets typically comprised only a few genetic regions from several dozen individuals, whereas modern datasets often contain hundreds or thousands of regions from hundreds of individuals. Increasing number of taxa and number of loci each pose their own difficulties. While there are 105 possible rooted, bifurcating trees with 5 taxa, there are about 1.3×10^{25} trees with 22 taxa (Felsenstein 1978) and many, many more with 100 taxa. Hence, with next-generation datasets, it is not feasible to examine every topology in tree space, since computational complexity increases rapidly with increasing number of taxa. More genetic regions also increase computational complexity, but in a different way.

Phylogenetic trees represent hypotheses about the evolutionary history of a set of taxa. Inferring a species tree given a single gene tree is straightforward, but integrating the signal between multiple conflicting gene trees is computationally difficult. Software such as *BEAST (Heled and Drummond 2010) can estimate a species tree from multiple gene trees in a Bayesian framework. An alternative approach, BUCKy (Larget 2010), measures the concordance between gene trees to summarize support for different species trees. Neither method can handle large modern phylogenetic datasets, so phylogeneticists often resort to concatenation, which forces each genetic region to have the same gene tree. While this approach is tractable, it can suffer from systematic error. When mathematical models do not account for important processes, they may result in systematic error. In the

concatenation approach, the assumption that all genes have the same underlying topology ignores relevant biological processes such as incomplete lineage sorting, which can cause gene trees with the same species tree to have different topologies. Problematically, the issue of systematic error is exacerbated by the addition of more data, since it may lead to increased confidence in erroneous conclusions.

Minimizing error and false confidence are important goals in any statistical discipline, but are current challenges in phylogenetics due to the massive influx in data and the lack of methods to interpret it. This thesis concerns the development and application of statistical methods to test for systematic error and minimize it when possible.

Chapter 1 provides an overview of some of the major current challenges in statistical phylogenetics.

The second chapter involves the use of informed priors to minimize error in Bayesian estimates of tree length. Previous studies have found that tree-length credible intervals in Bayesian analyses often exceed maximum likelihood estimates by an order of magnitude. These differences have been attributed to overly informative branch-length priors. I develop software to acquire similar datasets, estimate priors from those datasets, and then parameterize priors for the focal dataset. By itself, this approach greatly improves upon default priors and, when combined with previously developed tree-length distributions, eliminates the inflated tree length problem for most exemplar datasets.

Chapter 3 seeks to identify if systematic error could be the source of discordance between morphological and molecular squamate datasets. I refine existing software which uses posterior predictive simulation to investigate model adequacy for the molecular dataset, but fail to find evidence of systematic error.

Chapter 4 summarizes the results of chapters 2 and 3 while recommending directions for future work.

Chapter 2

Deflating Trees: Improving Bayesian Branch-length Estimates using Informed Priors

Setting priors is a necessary step in any Bayesian analysis, but the best approach to choice of priors has been a contentious issue in phylogenetics as it has in many other fields of statistical inference (Efron 2013). Approaches vary widely and different priors have been shown to influence the results of Bayesian phylogenetic analyses (Brown et al. 2010, Marshall 2010, Rannala et al. 2012, Nowak et al. 2013). Currently, choice of priors in phylogenetics is often justified by appeals to objectivity or robustness, including most of the default distributions used in popular software packages (MrBayes; Ronquist et al. 2012, BEAST; Drummond et al. 2012). Here, we consider a phylogenetic problem (branch-length inference) where the default prior approach can give rise to misleading conclusions and propose a solution using informed priors. Our informed prior approach leverages the information in previously published data to set priors for new analyses.

Some researchers favor the use of reference priors, which are selected by formal rules (Kass and Wasserman 1996) to represent a lack of information about the distribution of a particular parameter and are motivated by the desire to minimally influence the posterior distribution, so that the data determine support for each hypothesis (Gelman et al. 2003). In practice, finding reference priors is often difficult. While they are noninformative for the parameter of interest, they may induce a highly informative, implicit prior on some other parameter in ways that can be difficult to predict. Whether or not reference priors can be set legitimately has been a source of controversy in statistics for centuries (Efron 2013).

Another approach to setting priors, empirical Bayes, parameterizes the prior using the focal data (Efron 2013). Empirical Bayes has the advantage of guaranteeing that parameter values near the peak of the likelihood surface will have high prior weight, but has been criticized as non-Bayesian because the priors are dependent on the focal data and lead to artificially reduced estimates of uncertainty. Consequently, many Bayesians oppose its use (e.g. Rannala et al. 2012).

If no explicit statistical framework is embraced for choice of priors, software developers often set default values for a prior which work well for datasets on which they have been tested. This approach is practical and may be effective for most analyses, but it comes with no guarantees. Critically, users may not be aware when the default is unreasonable for their data, which can lead to the publication of erroneous conclusions. As a result, some phylogenetic software packages force users to manually set priors for important parameters (e.g. BEAST; Drummond et al. 2012).

In contrast to the above methods, informed priors make use of previous analyses, similar datasets, or expert opinion to set priors for the focal data, and have been used occasionally in phylogenetics (e.g. Liang et al. 2009, Goldstein 2006, Nowak et al. 2013). Informed priors are, confusingly, not necessarily informativethe former are set using outside information, whereas the latter means that the prior strongly influences the posterior. Informed priors have the advantage of incorporating current knowledge directly into the model, which typically leads to tighter confidence intervals around parameter estimates than the reference prior approach.

The informed prior approach is not without its pitfalls. While phylogenetic databases from which relevant information could be extracted do exist (e.g. TreeBase, http://www.treebase.org; DRYAD, http://datadryad.org), they are difficult

to query, which makes finding relevant datasets difficult. Alternatively, for priors that should be set to reflect expert opinion, it can be difficult to quantify that opinion into a range of relevant parameter values, particularly for phylogenetic analyses where wide-ranging factors such as divergence time, sampling structure, and choice of loci have a strong influence on conclusions.

2.1 Branch-length Priors in Bayesian Phylogenetics

To illustrate the importance of effective prior choice in phylogenetics, we consider the problem of branch length inference. Branch-length estimates are often of direct interest in phylogenetic analyses, since they describe the amount of evolutionary change between nodes. These estimates can affect a wide variety of biological inferences including ancestral state reconstruction, species delimitation, divergence time estimation, and rates of lineage diversification. Hence, it is concerning that Bayesian estimates of total tree length can be an order of magnitude longer than maximum likelihood (ML) estimates (Brown et al. 2010, Marshall 2010) when alignments contain many closely related sequences and default branch-length priors from standard software packages (e.g., MrBayes; Ronquist et al. 2012) are used.

Brown et al. (2010) investigated the issue of inflated tree lengths in a range of simulated and empirical datasets, examining whether the problem may be due to mixing problems for the Markov chain caused by (1) multiple local peaks or (2) large, nearly flat regions in the posterior. Alternatively, (3) an overly informative branch-length prior may bias the posterior towards unreasonably large branch lengths. They found support for possibilities (2) and (3). Further analysis by Rannala et al. (2012) suggested that a poorly specified branch-length prior could be the root cause of all three possibilities.

By default, MrBayes uses independent and identically distributed exponential priors with a rate (λ) of 10 for branch lengths (Ronquist and Huelsenbeck

2003, Ronquist et al. 2012). For a tree with n taxa, total tree length is the convolution of $2n-3$ exponential densities, which has the gamma distribution with shape $\alpha = 2n-3$ and rate λ (Fig. 2.1). Since the mean of the gamma distribution is α/λ, expected tree length is $(2n-3)/\lambda$. Hence, the branch-length prior sets an implicit prior on tree length that scales with the number of taxa and can be highly sensitive to changes in λ (see figure 7 from Brown et al. 2010). For phylogenies with recent divergences and many taxa, this sensitivity often leads to default prior tree-length distributions that exclude the ML estimate (Fig. 2.1).

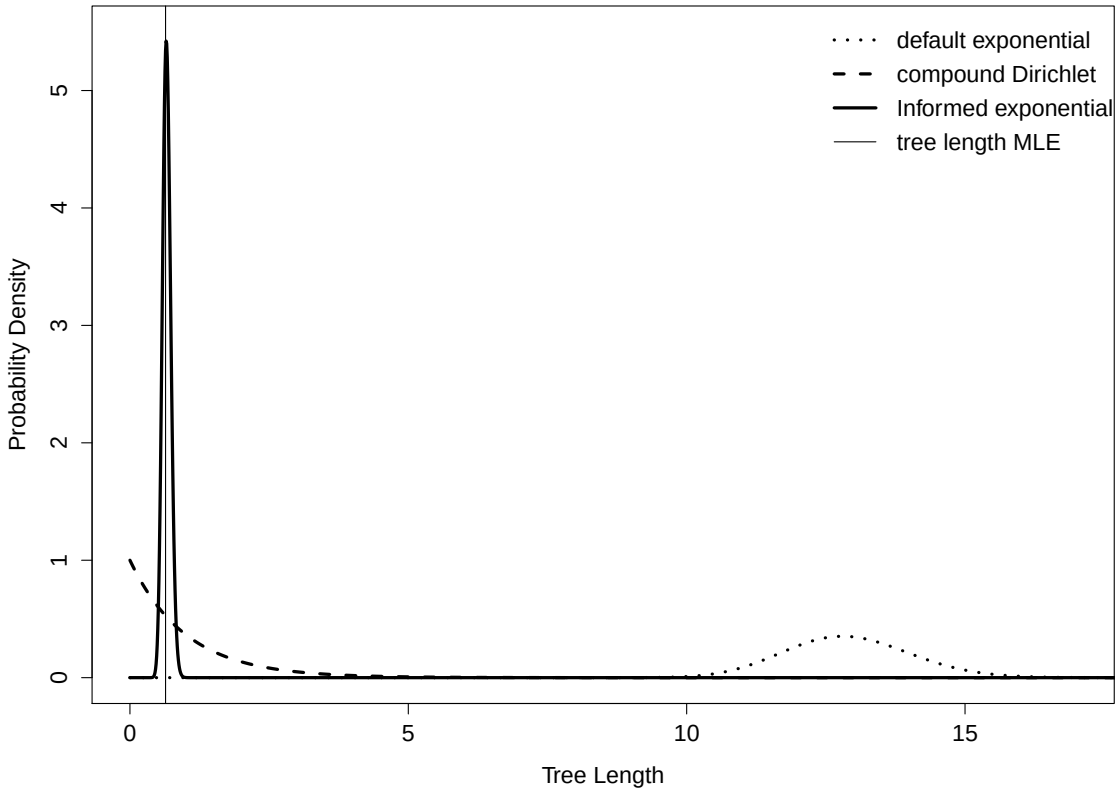

FIGURE 2.1. Comparison of tree-length prior distributions for default and informed exponential branch-length priors, as well as the default compound Dirichlet tree-length prior, to the MLE for *Acris* tree length. The informed exponential was parameterized using TreeBASE dataset S10170 obtained from EmpPrior.

Several approaches have been proposed to mitigate the influence of overly informative default branch-length priors. Brown et al. (2010) recommended an empirical Bayes approach, which recovered the ML tree-length estimate in credible intervals, but suffers from an artificial reduction in uncertainty. Other approaches have aimed to set less informative default priors on branch- or tree-length, including the double-exponential branch-length prior (Yang and Rannala 2005) and the compound Dirichlet tree-length prior (Fig. 2.1; Rannala et al. 2012, Zhang et al. 2012), both of which are implemented (but not used by default) in MrBayes 3.2. The double-exponential branch-length prior specifies separate exponential priors on internal and external branches, whereas the compound Dirichlet prior sets a (inverse) gamma-distributed prior on total tree length, a Dirichlet prior on the partitioning of tree length among branches, and a beta-distributed prior on the partitioning of tree length between internal and external branches. The compound gamma Dirichlet tree-length prior (as implemented by Zhang et al. [2012] in MrBayes) has four parameters: tree length shape (α_T) and rate (λ_T), Dirichlet concentration (α), and mean internal:external branch length ratio (c) with a default value of 1 for all parameters. The double-exponential branch-length prior marginally shrinks tree-length estimates but is still highly sensitive to prior settings, whereas the compound Dirichlet prior successfully recovers ML total tree length estimates in several (but not all) problematic datasets across a wide range of tree length prior means (Zhang et al. 2012).

Here, we propose an extension to the default prior approaches mentioned above that involves setting informed priors based on outside data. We compare posteriors from informed priors to default priors across a range of datasets that have produced problematic branch-length estimates (Brown et al. 2010, Zhang et al. 2012). We investigate both exponential branch-length priors and recently pro-

posed compound Dirichlet tree-length priors (Rannala et al. 2012). As we show, informed priors greatly improve upon default priors and produce 95% highest posterior density (HPD) intervals that typically include ML estimates for both the exponential and compound Dirichlet distributions. Given the wide availability of outside phylogenetic data, we recommend increased use of informed priors in Bayesian phylogenetic analyses.

2.2 Obtaining Informed Prior Estimates from Phylogenetic Databases

In order to obtain informed prior estimates, we must acquire datasets that are relevant to our focal data. We used 3 criteria to establish the relevance of a particular external dataset. Relative to the focal data, external datasets should (1) include orthologous regions of DNA, (2) have a similar number of taxa, and (3) sample taxa with a similar degree of divergence. Properties (1) and (2) are relatively easy to test, but it may be difficult to confirm (3) without estimating parameters from the focal data. To circumvent the issue, we used taxonomic classification as a rough proxy for divergence, using external datasets only if they had similar taxonomic depth, number of species, and number of samples per species as the focal dataset. Taxonomic classification is not necessarily strongly correlated with divergence, but should provide a rough approximation of tree depth and set a useful upper bound on divergence.

We developed a program, EmpPrior (Available from code.google.com/p/empprior/; Andersen et al. 2014), to parameterize branch- and tree-length distributions by searching TreeBase for datasets similar to focal data. EmpPrior comprises a Java program, EmpPrior-search, which queries TreeBase and returns matching datasets, and an R script, EmpPrior-fit, which finds ML parameter estimates for exponential branch-length and compound gamma Dirichlet tree-length distributions. We perform ML tree search on each dataset returned from TreeBase, infer parameter

estimates for branch- and tree-length distributions from resulting topologies, and inform focal priors based on these estimates. While the use of ML methods in a Bayesian analysis may seem unusual, it is justified in this case as a fast approximation that integrates easily into current software and consistently outperforms default settings (see below). To facilitate comparison of inferences across studies, we used exemplar empirical datasets previously analyzed by Brown et al. (2010) and later used by Zhang et al. (2012) to test alternative branch- and tree-length prior distributions. The four chosen datasets represent a diverse set of animal clades with shallow divergences and many sequences: *Acris* (cricket frog, 66 sequences, Gamble et al. 2008), *Corbicula* (freshwater clam, 93 sequences, Hedtke et al. 2008), *Crinia signifera* (common eastern froglet, 92 sequences, Symula et al. 2008), and *Sceloporus* (spiny lizard, 123 sequences, Leaché and Mulcahy 2007).

For each focal dataset, we used EmpPrior to search TreeBase for datasets with the same regions of DNA and a similar number of taxa (using a cutoff of ± 20 taxa). Retrieved datasets with similar taxon sampling were used in downstream analyses. For each retrieved dataset, we estimated ML trees using Garli v2.0 (Zwickl 2006). We used the Nelder-Mead method (Nelder and Mead 1965) as implemented in the optim function in R 3.0 (R Core Team 2013) and the R package bbmle (Bolker 2013) to perform maximum-likelihood estimation of parameters for exponential branch-length distributions and compound Dirichlet tree-length distributions. In addition, we compared 8 submodels of the compound Dirichlet model where λ_T, α, and c were either fixed at default values or estimated from the data. We set $\alpha_T = 1$ in all cases to represent a diffuse prior on tree length and because joint estimation of α_T and λ_T was unreliable.

We used MrBayes v3.2.1 (Ronquist et al. 2012) as modified by Zhang et al. (2012; includes compound Dirichlet tree-length prior) for Bayesian phylogenetic

analyses. Each focal dataset was analyzed using default and informed exponential priors, as well as default and informed compound Dirichlet priors. Each analysis was run for 5,000,000 generations with two independent runs and four chains per run, 25% burn-in and samples recorded every 1000 generations. Convergence was assessed using the average standard deviation of split frequencies ($ASDSF < 0.01$; Lackner et al. 2008) and trace plots in Tracer v1.5 (Rambaut and Drummond 2009). We used R package ggplot2 (Wickham 2009) to create violin plots of posterior density. ML estimates of tree length from each focal dataset were obtained from Brown et al. (2010).

2.3 Empirical Performance of Default and Informed Prior

The *Acris* dataset contained 66 *cyt b* sequences from two species. The EmpPrior search returned 24 nexus files, 10 of which included intra-generic sampling. From the intra-generic datasets, 3 datasets that included 10 or more sequences for multiple species were selected for further analysis. The default exponential branch-length prior did not recover the ML tree length estimate (95% CI: [0.725, 1.302], MLE: 0.64, Table 2.1), whereas two of the three informed exponential priors recovered the MLE in 95% credible intervals (CIs) and the third barely excluded it (Fig. 2.2a, Table 2.1). Both informed and default compound Dirichlet priors recovered the ML estimate (Fig. 2.2b, Table 2.1). Informed compound Dirichlet priors resulted in CIs with medians substantially closer to ML values than compound Dirichlet priors with default values (Fig. 2.2b, Table 2.1). We focus here primarily on the effects of using compound Dirichlet priors with informed values of α (Fig. 2.2), because these models tended to produce narrow credible intervals that still included ML tree lengths from the focal data sets. Full results for compound Dirichlet priors with informed values of c, or a combination of α and c are given in Table 2.1.

The *Corbicula* dataset comprised 93 *COI* sequences from over 8 species. The EmpPrior search returned 44 nexus files, 18 of them at genus level. Two datasets that included multiple sequences for at least 7 species were selected. The default exponential branch-length prior did not recover the ML tree-length estimate (95% CI: [9.254, 20.861], MLE: 1.96, Table 2.1), whereas both of the informed priors recovered the MLE in 95% CIs (Fig. 2.2c, Table 2.1). The default compound Dirichlet prior did not recover the MLE and, while only one of the informed priors recovered the MLE, both of them improved upon default settings (Fig. 2.2d, Table 2.1).

The *Crinia signifera* dataset contained 92 concatenated *12S* and *16S* sequences from a single species. EmpPrior found 12 nexus files that contained both genes, none of which were entirely intraspecific. However, intraspecific sequences were extracted from two datasets to create datasets with 60 and 48 intraspecific sequences. The default exponential branch-length prior did not recover the ML tree-length estimate (95% CI: [1.189, 4.510], MLE: 0.55, Table 2.1), whereas one of the two informed priors recovered the MLE in 95% CIs (Fig. 2.2e, Table 2.1). Both the informed and the default compound Dirichlet priors recovered the MLE, with the informed CI medians closer to the MLE than the default (Fig. 2.2f, Table 2.1).

The *Sceloporus* dataset contained 123 sequences for two genes, *nad4* and *12S*, with multiple sequences for each of 14 species. EmpPrior found a single dataset for each gene, which yielded over 20-fold differences in estimates using EmpPrior-fit (9 and 210, respectively). Neither informed nor default exponential branch-length priors recovered the ML tree length, but one informed estimate greatly shrank both the median and the width of the credible interval (Fig. 2.2g, Table 2.1). One of the informed compound Dirichlet estimates yielded a CI that included the ML

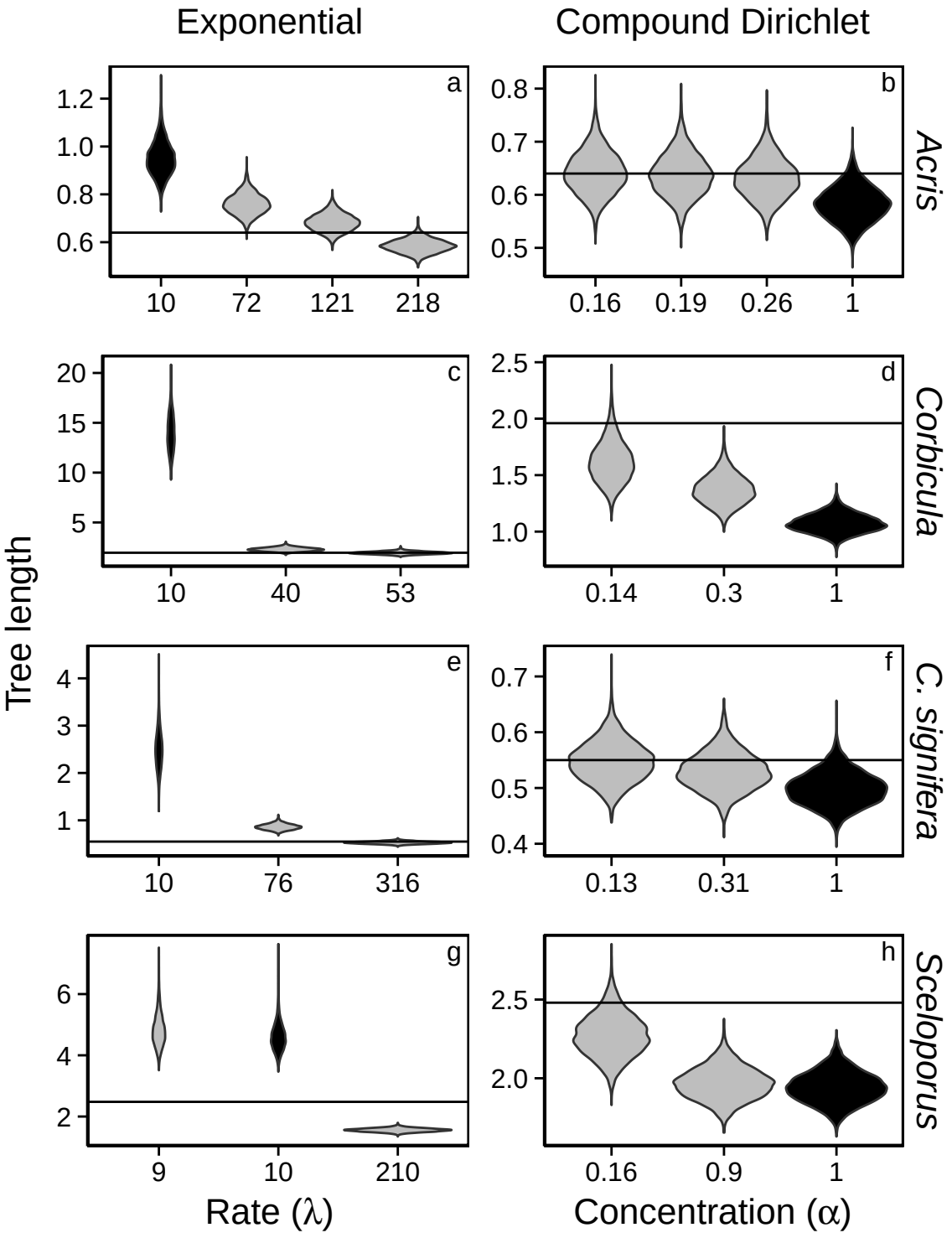

FIGURE 2.2. Violin plots of tree length (TL) posterior density for a-b) *Acris*, c-d) *Corbicula*, e-f) *Crinia signifera*, and g-h) *Sceloporus*. Plots resulting from default priors have a black fill whereas those from informed priors have a grey fill. ML TL estimates are indicated with a solid horizontal line. Results from exponential priors are in the left column (a, c, e, and g) and results from compound Dirichlet priors are in the right column (b, d, f, and h).

tree length, whereas the other informed and the default CIs fell below the ML tree length (Fig. 2.2h, Table 2.1).

For the exponential branch-length prior, the informed approach dramatically improved tree-length credible intervals relative to MrBayes defaults in three of the analyzed datasets. The majority (2/3 for *Acris*, 2/2 for *Corbicula*, 1/2 for *Crinia signifera*, 0/2 for *Sceloporus*) of analyses using informed priors recovered the ML tree-length estimate in credible intervals, whereas tree-length estimates from the default exponential prior were an order of magnitude too large. The improvement in tree-length estimates for informed priors suggests that taxonomic classification was a reasonable proxy for sequence divergence for our focal datasets, but more direct measures of divergence might prove more reliable.

For the compound Dirichlet tree-length prior, the majority of informed prior CIs included ML tree lengths (3/3 for *Acris*, 1/2 for *Corbicula*, 2/2 for *Crinia signifera*, 1/2 for *Sceloporus*). In addition, CIs resulting from the default compound Dirichlet tree-length prior often included ML tree-length estimates (2/4 datasets), but informed CI medians were often substantially closer to ML tree-length estimates. Overall, using informed estimates of α improved upon default settings for *Corbicula* and *Sceloporus* (where the default did poorly) and never led to worse estimates.

Estimates of Dirichlet concentration (α) were generally less than 1, which changes the shape of the distribution from flat (at $\alpha = 1$) to U-shaped, with more prior weight on both large and small relative branch lengths. This distribution seems reasonable for the exemplar datasets, which have many short intraspecific branches and a few longer interspecific ones. Estimates of c were generally greater than 1, meaning that internal branches were, on average, longer than external ones. This result contradicts the expectation of Rannala et al. (2012) that mean

internal:external branch-length ratio should generally be less than 1. However, these estimates make sense for our focal datasets, which include many small, intraspecific terminal branches. This result illustrates the utility of using empirical estimates to set informative priors, since a rule of thumb that makes sense for deep divergences may not be reasonable for shallow divergences.

When inferring trees for use with informed priors, algorithmic approaches such as neighbor joining (NJ) and approximate ML software such as phangorn or FastTree 2 (Price et al. 2010) consistently yielded shorter trees than full ML approaches such as Garli (Zwickl 2006). These shorter trees resulted in larger exponential rate estimates and smaller posterior mean tree lengths that often failed to include the ML estimate in 95% CIs. This downward bias may be due to difficulties inherent in estimating many short branches in retrieved datasets, since the NJ method in ape (Paradis et al. 2004) often returned negative branch lengths. However, setting a positive minimum bound on NJ branch-length estimates did not fix the issue. We recommend using a full ML search to inform rate estimates for the exponential branch-length prior.

The *Sceloporus* dataset presents an interesting challenge for the application of informed priors, since averaging widely divergent rate estimates could be error-prone. EmpPrior found few datasets that contained both *12S* and *ND4* sequences, which may have led to use of non-relevant datasets in our analysis. In particular, the external *ND4* dataset is suspect because it lacks intraspecific sampling, which may have contributed to its low exponential rate estimate. In cases where different genes have large differences in rate, setting separate informed branch-length priors for separate genes or scaling priors based on gene may be appropriate.

TABLE 2.1: EmpPrior datasets and models for exponential and compound Dirichlet models with 95% credible intervals. TL CIs that include the MLE are indicated in bold. 'Focal' refers to the empirical dataset, 'ID' refers to the TreeBASE ID of the dataset returned by EmpPrior, 'EX' means exponential, and 'CD' means compound Dirichlet with suffixes a (concentration) and c (branch-length ratio) used to indicate which parameters of the compound Dirichlet are being fitted. LL is the log-likelihood, α_T and λ_T are the parameters for the gamma tree-length distribution, α refers to concentration, c is the branch-length ratio, and MLE TL is the maximum-likelihood estimate for total tree length. Rows without a TreeBASE ID correspond to default MrBayes settings.

Focal	**ID**	**Model**	**LL**	**df**	**dAICc**	**weight**	**Mean TL**	α_T	λ_T	α	c	**95% TL CI**	**MLE TL**
Acris	S10170	CD.ac	-445.831	2	0	0.999	1	1	1	0.153	2.624	**[0.509, 0.797]**	0.64
Acris	S10170	CD.a	-437.759	1	14.04	0.001	1	1	1	0.189	1	**[0.500, 0.801]**	0.64
Acris	S10170	EX	-303.813	1	281.93	0	0.66	80	121.227	1	1	**[0.567, 0.819]**	0.64
Acris	S10170	CD.c	-271.054	1	347.448	0	1	1	1	1	0.709	**[0.531, 7.984]**	0.64
Acris	S1800	CD.ac	-428.578	2	0	1	1	1	1	0.211	2.74	**[0.491, 0.791]**	0.64
Acris	S1800	CD.a	-418.237	1	18.59	0	1	1	1	0.259	1	**[0.515, 0.797]**	0.64
Acris	S1800	EX	-394.706	1	65.652	0	0.412	90	218.247	1	1	**[0.494, 0.706]**	0.64
Acris	S1800	CD.c	-313.27	1	228.525	0	1	1	1	1	1.038	**[0.541, 4.041]**	0.64
Acris	S12419	CD.ac	-587.909	2	0	1	1	1	1	0.123	2.836	**[0.494, 0.795]**	0.64
Acris	S12419	CD.a	-576.998	1	19.732	0	1	1	1	0.158	1	**[0.507, 0.827]**	0.64
Acris	S12419	CD.c	-323.799	1	526.129	0	1	1	1	1	0.596	**[0.532, 8.358]**	0.64
Acris	S12419	EX	-298.198	1	577.331	0	1.264	91	72.017	1	1	**[0.612, 0.957]**	0.64
Acris		EX							10			[0.725, 1.302]	0.64
Acris		CD					1	1	1	1	1	**[0.462, 0.728]**	0.64
Corbicula	S10579	CD.ac	-770.181	2	0	1	1	1	1	0.107	4.718	**[0.913, 2.019]**	1.96
Corbicula	S10579	CD.a	-747.002	1	44.279	0	1	1	1	0.138	1	**[1.095, 2.480]**	1.96
Corbicula	S10579	CD.c	-375.719	1	786.844	0	1	1	1	1	1.046	**[1.109, 3.380]**	1.96
Corbicula	S10579	EX	-309.852	1	918.578	0	1.945	104	53.482	1	1	**[1.511, 2.639]**	1.96
Corbicula	S1910	CD.ac	-343.481	2	0	0.998	1	1	1	0.233	2.294	[0.913, 1.210]	1.96
Corbicula	S1910	CD.a	-336.201	1	12.456	0.002	1	1	1	0.3	1	[0.994, 1.938]	1.96
Corbicula	S1910	CD.c	-267.887	1	149.083	0	1	1	1	1	0.849	**[1.240, 4.818]**	1.96
Corbicula	S1910	EX	-214.187	1	256.484	0	2.023	80	39.541	1	1	**[1.725, 3.083]**	1.96
Corbicula		EX							10			[9.254, 20.861]	1.96
Corbicula		CD					1	1	1	1	1	[0.771, 1.427]	1.96
Csignifera	S10211	CD.ac	-2579.913	2	0	1	1	1	1	0.248	2.369	**[0.407, 0.646]**	0.55
Csignifera	S10211	CD.a	-2540.097	1	77.613	0	1	1	1	0.308	1	**[0.411, 0.660]**	0.55
Csignifera	S10211	CD.c	-2192.31	1	773.187	0	1	1	1	1	0.962	[0.552, 3.595]	0.55
Csignifera	S10211	EX	-2059.185	1	1039.437	0	1.37	433	315.963	1	1	**[0.438, 0.623]**	0.55
Csignifera	S13567	CD.ac	-3334.805	2	0	1	1	1	1	0.095	5.546	**[0.395, 0.660]**	0.55
Csignifera	S13567	CD.a	-3231.365	1	204.859	0	1	1	1	0.13	1	**[0.438, 0.740]**	0.55

(Table 2.1 continued)

Focal	**ID**	**Model**	**LL**	**df**	**dAICc**	**weight**	**Mean TL**	α_T	λ_T	α	c	**95% TL CI**	**MLE TL**
Csignifera	S13567	CD.c	-1805.499	1	3056.59	0	1	1	1	1	1.06	**[0.528, 3.821]**	0.55
Csignifera	S13567	EX	-1228.928	1	4209.732	0	4.857	369	75.977	1	1	[0.676, 1.120]	0.55
Csignifera		EX							10			[1.189, 4.510]	0.55
Csignifera		CD					1	1	1	1	1	**[0.394, 0.658]**	0.55
Sceloporus	S10211	CD.ac	-1328.424	2	0	1	1	1	1	0.128	5.36	**[1.687, 2.611]**	2.48
Sceloporus	S10211	CD.a	-1283.967	1	86.871	0	1	1	1	0.159	1	[1.652, 2.381]	2.48
Sceloporus	S10211	EX	-796.019	1	1062.765	0	0.869	183	210.573	1	1	[3.493, 7.537]	2.48
Sceloporus	S10211	CD.c	-775.071	1	1104.661	0	1	1	1	1	1.554	**[2.290, 7.490]**	2.48
Sceloporus	S10106	CD.ac	-394.515	2	0	0.517	1	1	1	1.245	0.603	[1.636, 2.413]	2.48
Sceloporus	S10106	CD.c	-393.396	1	0.161	0.477	1	1	1	1	0.688	[2.582, 14.700]	2.48
Sceloporus	S10106	CD.a	-389.04	1	8.874	0.006	1	1	1	0.897	1	**[1.828, 2.854]**	2.48
Sceloporus	S10106	EX	-128.525	1	529.903	0	12.332	109	8.839	1	1	[1.345, 1.796]	2.48
Sceloporus		EX							10			[3.452, 7.645]	2.48
Sceloporus		CD					1	1	1	1	1	[1.629, 2.306]	2.48

2.4 Relevance of Informed Priors in Bayesian Phylogenetics

Informed branch-length priors obtained with EmpPrior often dramatically improve upon default prior settings and at worst seem to cause no harm. Exploring additional applications of informed priors in phylogenetics may thus prove fruitful. In particular, using previous datasets to inform priors on rate variation may improve estimates of divergence times.

While we have shown that informed branch-length priors often improve tree-length estimates for datasets with shallow divergence and many taxa, we have not tested the method on phylogenies with deeper divergences. Since the inflated tree problem seems specific to datasets with shallow divergences, improvements to branch-length estimates using the informed approach may be greatest for these datasets. The effectiveness of informed priors in other circumstances remains an open question.

As publicly available phylogenetic databases become larger and simpler to query, obtaining outside information to parameterize priors will become easier and these estimates may also become more accurate. EmpPrior provides a simple and effective way to query an existing database (TreeBase) for similar datasets and obtain informed branch-length prior parameters. Since using informed priors can deflate Bayesian trees that have grown too long, informed priors should be part of every phylogeneticists toolkit.

Chapter 3

Posterior Prediction in Squamate Phylogenetics

When phylogenies inferred from different datasets are incongruent, reconciling them can be difficult. These conflicts are not uncommon in evolutionary biology, where inferences based on molecular data can contrast sharply with traditional taxonomies and conclusions from morphological data. Sometimes, conflicts are resolved with additional data (e.g. Gatesy and OLeary 2001, Naylor and Adams 2001), but systematic biases caused by implausible models have also been identified as a source of error in molecular phylogenies (DErchia et al. 1996, Sullivan and Swofford 1997). Recently, Wiens et al. (2012) and Gauthier et al. (2012) analyzed what are, respectively, the largest molecular and morphological squamate datasets ever assembled, and came to divergent but strongly supported conclusions regarding the evolutionary history of snakes and lizards. In particular, the molecular tree placed dibamids and gekkotans sister to all other squamates, whereas the morphological tree, in accordance with traditional taxonomic placement, strongly supported an Iguania-Scleroglossa basal split, a Gekkota-Autarchglossa sister-group, and the divergence between Anguimorpha and Scinomorpha (Gauthier et al. 2012). Losos et al. (2012) examined the implications of the molecular phylogeny for the morphological data, concluding that the implied number of morphological reversals was implausible, particularly for Iguania. However, the morphological data were not analyzed in a statistical framework, making direct model comparisons difficult.

While much effort has been devoted to weighing the relative merits of molecular and morphological data (e.g. Hillis 1987, Swofford 1991, Wendell and Doyle 1998, Losos et al. 2012), here we focus on the adequacy of the phylogenetic models used to explain the molecular data. Many sources of error can affect phylogenetic models, such as simplistic models of nucleotide evolution (Sullivan and Swofford 1997), but systematic error in general is particularly problematic because it can result in high confidence in a poor model, a problem which may be exacerbated by additional data (Naylor and Brown 1997).

In a Bayesian framework, it is assumed that the true model is part of the family of models under consideration. This assumption is almost always violated in phylogenetic analyses, since most models do not account for complex evolutionary processes such as (list examples). The degree to which this model violation affects inference is often unclear, but several studies have found strong effects of model violation on phylogenetic conclusions (Swafford and Sullivan 1997, Brown et al. 2010, Brown 2014).

Several approaches have been developed to test for model violation in a Bayesian framework, including posterior prediction (Bollback 2002, Waddell et al. 2009, Brown 2014). Posterior prediction is an important tool for assessing the ade-

quacy of Bayesian models (Gelman et al. 2013) which, despite being applied to phylogenetics over a decade ago (Bollback et al. 2002), has not been widely adopted. In typical posterior prediction analyses, new datasets are simulated using parameters drawn from the posterior distribution of an empirical analysis. Test statistics are computed from these simulated datasets, creating a distribution which can be compared to the empirical test statistic value. If the empirical value falls outside the distribution, then some degree of model violation may be occurring. Importantly, failure to detect model inadequacy does not imply that there is no model violation, since the model may be inadequate in some way that is not revealed by the summary statistics. Inferential test statistics (Brown et al. 2014) expand the traditional posterior predictive approach in phylogenetics by generating tree distributions from simulated data using software such as MrBayes (Ronquist and Huelsenbeck 2003). These new data allow for a slew of phylogenetically-relevant test statistics such as mean tree length, RF-distance quantiles, and topological entropy, and have detected model violation in exemplar simulated and empirical datasets (Brown et al. 2014), but have not been widely tested.

Here, our focus is to 1) assess model adequacy in the squamate molecular dataset and discuss how model violations (if any) affect phylogenetic inference and 2) to assess the performance of various inferential test statistics across a large dataset.

3.1 Methods

We used the squamate molecular dataset developed by Wiens et al. 2012 and, for computational feasibility, chose the 31-taxa subsample they used for *Beast analysis (Heled and Drummond 2010). These data comprised 44 genes and represented all major squamate families. We performed empirical analyses on each gene separately using the GTR+γ model in MrBayes v3.2.2 (Ronquist and Huelsenbeck 2003), sampling every 1000 generations for 4 million generations, with each gene partitioned by codon position. This matches the analyses performed by Wiens et al. (2012), with the exception that they concatenated their genes. Convergence was assessed by ensuring average standard deviation of split frequencies was ¡0.01. Trace plots were inspected manually in TRACER (Rambaut and Drummond 2009), and optimal burn-in was assessed by MrConverge v1b2 (written by A.R. Lemmon). For each gene, 100 posterior predictive datasets were simulated using uniform draws from the posterior distribution (post-burnin) in PuMA (Brown and ElDabaje 2009). Phylogenetic analyses on the posterior predictive datasets were identical to the empirical, with the exception that MrBayes analyses were run for 3 million generations and a default burn-in of 0.25 was used in place of MrConverge estimates.

We calculated 10 summary statistics using AMP (Brown 2014), including topological entropy, tree length (TL) mean, TL variance, weighted Robinson-Foulds (RF; Robinson and Foulds 1981) distance interquartile range (IQR), and 1-4, 2-4, 3-4, 99-100, 999-1000, and 9999-10000 RF quantiles, as well as effect sizes and two-tailed p-values. In addition, we calculated symmetric difference (Robinson and Foulds 1981) between the Wiens et al. (2012) topology and topologies from

the posterior distribution of each gene using Dendropy (Sukumaran and Holder 2010). We used matplotlib (Hunter 2007) to generate histograms and scatterplots in Python.

3.2 Results

The 99-100, 999-1000, and 9999-10000 RF quantile test statistics were variable in 13, 12, and 12 genes, respectively, with 12 variable genes in common. In contrast, the 1-4, 2-4, and 3-4 RF quantile test statistics were variable in 44, 44, and 42 genes, respectively. The entropy test statistic was invariant within individual genes, but the empirical entropy was always lower than the mean. Both TL mean and TL variance were variable within each gene. The distribution of mean symmetric differences across genes was bimodal (Fig. 1) and the minimum symmetric difference was 19, indicating that no posterior distribution of gene trees contained the Wiens topology. Mean symmetric difference between the Wiens topology and posterior gene topologies was positively correlated with RF distance IQR ($r = 0.35$, $p = 0.021$, Fig. 2a) and negatively correlated with both TL variance ($r = 0.39$, $p = 0.009$, Fig. 2b) and the RF 1-4 quantile ($r = -0.31$, $p = 0.044$, Fig. 2c). No other mean symmetric difference to summary statistic correlations were significant ($p > 0.05$). In general, genes with lower RF distances appeared to have summary statistic effect sizes closer to zero.

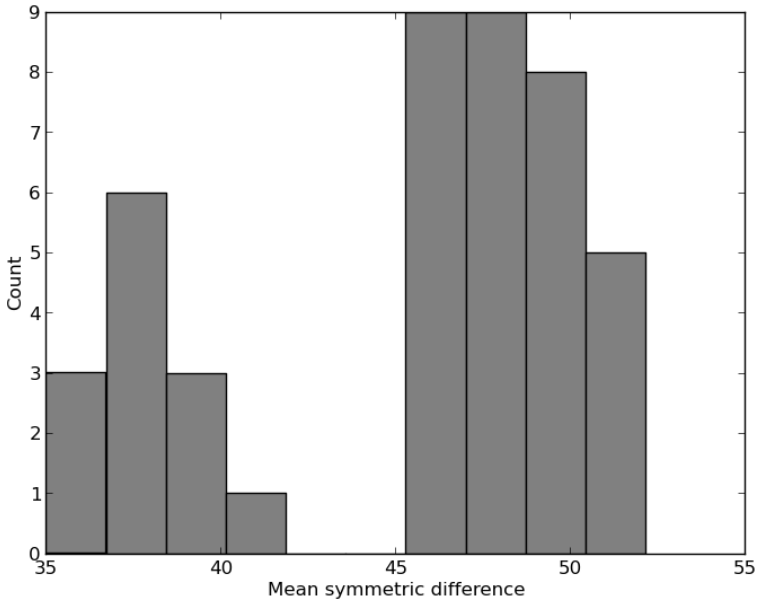

FIGURE 3.1. Histogram illustrating bimodal distribution of mean symmetric difference between the *BEAST topology in Wiens et al. 2012 and topologies from the posterior distribution from each gene.

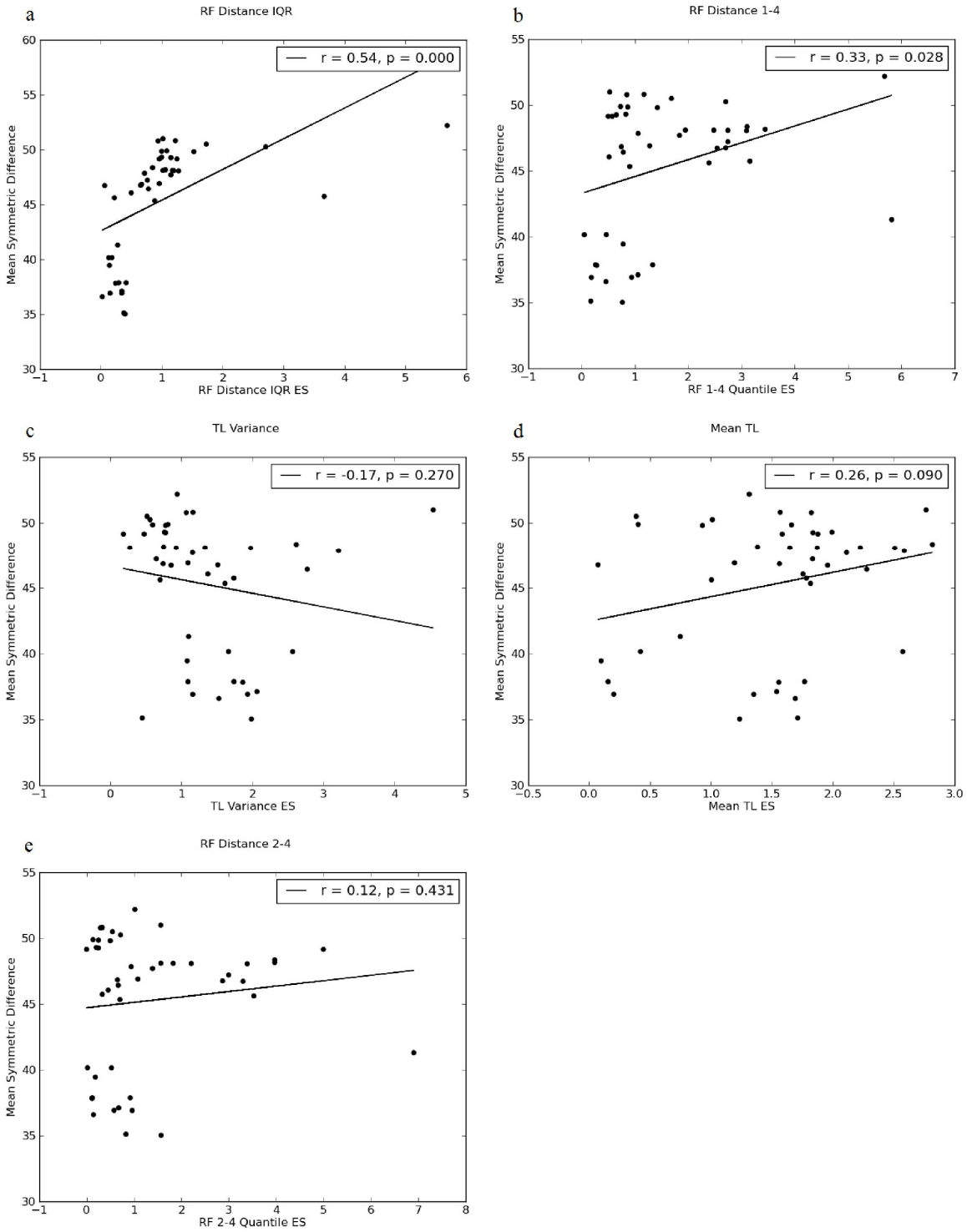

FIGURE 3.2. Scatterplots of mean symmetric difference versus effect size for the following posterior predictive summary statistics: a) RF IQR, b) TL variance , and c) RF 1-4 quantile. Lines of best fit are indicated in black, with correlation coefficient (r) and two-tailed p-value.

3.3 Discussion

Support for the Wiens topology among individual genes was both heterogeneous and relatively low. The bimodal distribution of symmetric differences suggests conflicting signal among genes. This could be due to a number of factors, including variation in phylogenetic signal strength or actual support for different topologies. The fact that the Wiens topology never appeared in the posterior distribution for any of the 44 genes does not suggest strong support for that topology, but the generally lower effect size magnitude among genes with smaller mean symmetric difference scores implies that genes with topologies closer to the Wiens topology had less model violation.

Several summary statistics were correlated with mean symmetric difference. The positive correlation between symmetric difference and RF IQR appears to have been strongly influenced by one outlier gene, but in general, genes with lower mean symmetric difference scores fall in the middle of their respective posterior predictive distributions, suggesting that genes that conflicted more strongly with the Wiens topology had RF IQRs that differed more from their posterior predictive distributions. The TL variance summary statistic was also positively correlated with mean symmetric difference and may have been influenced by an outlier gene. Genes with low (< 42) mean symmetric difference scores all had negative TL variance effect sizes. Lastly, the 1-4 RF distance quantile summary statistic was negatively correlated with mean symmetric difference, but was heavily influenced by two outlier genes which had much lower effect sizes. Most empirical datasets had negative effect sizes, indicating that their trees were more tightly clustered than posterior predictive datasets.

Summary statistic performance varied widely, and several (entropy; 99-100, 999-1000, and 9999-10000 RF quantiles) showed little or no variation among posterior predictive data sets within genes. The results of the entropy test statistic suggests that some topologies appeared more than once in all empirical analyses, but almost never in posterior predictive datasets. Sampling more trees from the posterior distribution in each posterior predictive dataset should eventually lead to within-gene variation in entropy scores, but the entropy test statistic is unlikely to perform well in datasets with more taxa.

There was a clear difference between quantiles near the center of the RF distance distribution (1-4, 2-4, and 3-4) and those near the tail (99-100, 999-1000, 9999-10000). Both sets might be improved by using matching distance (Bogdanowicz and Giaro 2012, Lin et al. 2012) instead of RF distance, since it has more reliable statistical properties.

Overall, we found evidence of model violation in many genes across many test statistics. However, the genes with the most model violation typically were least supportive of the Wiens topology. Hence, it is still unclear to what extent this systematic error influenced the Wiens analysis. Since different genes exhibited different degrees of model violation for different summary statistics, it can be difficult to determine which genes are most reliable. Future work could explore combining information from multiple test statistics using principle components

analysis or multiple regression and filtering genes which exhibit a high degree of model violation.

Conflicting signal among genes is a pattern routinely found in phylogenetics, and there is little reason to suspect it will be eliminated with additional loci. In order to make reliable inferences from this wealth of data, phylogeneticists should devote more time to model checking and validation, so that additional information does not lead to false confidence in the wrong topologies.

Chapter 4
Summary and Conclusions

The development of new statistical methods and the use of prior information can identify and reduce systematic error, but may not always reconcile discordant datasets. Informed priors are a promising approach that, when combined with tree-length priors, appear to offer a resolution to the problem of inflated Bayesian trees. In contrast, posterior predictive simulation failed to identify meaningful systematic error in the molecular squamate dataset, leaving the discordance between molecular and morphological datasets unresolved for the clade. Further refinement of these methods may result in better detection of model inadequacy and less error overall.

Repurposing of existing data for new analyses is common in science, but has not been used often to generate informed priors in phylogenetics. This could be attributed to difficulty finding datasets similar to the focal dataset, which may result from difficulty querying phylogenetic databases or the more general issue of what makes phylogenetic datasets similar. We find that simple filters such as genetic region, number of taxa, and tree depth can yield relevantly similar datasets, and that parameters estimated from these datasets greatly improve upon defaults in software such as MrBayes. When coupled with previous approaches such as the use of compound Dirichlet tree-length priors instead of exponential branch-length priors, the problem of inflated Bayesian tree-length estimates is greatly diminished in exemplar datasets and, in most cases, is entirely eliminated.

The general application of the informed prior approach to the tree-length problem in phylogenetics may be limited by the use of unnamed genetic regions in phylogenetic analyses, but the issue could be ameliorated by the development of more advanced phylogenetic databases that include metadata such as the location of genetic regions on reference chromosomes. A similar method could also be used to inform priors on rate variation for divergence time estimation.

While the posterior prediction methods failed to detect model inadequacy in the molecular squamate dataset, it is unclear if this is due to adequacy of the phylogenetic model or insufficient posterior predictive tests. These methods were previously tested primarily on simulated datasets with fewer taxa, so many of the posterior predictive tests were not tuned for these data. In particular, analyses on the simulated datasets found that quantiles on the right tail of the RF distance distribution (e.g. 99-100 quantile) were more sensitive to model inadequacy, whereas in the squamate dataset, quantiles on the left tail or center of the RF distance distribution appeared more sensitive.

In addition to tuning tests to the focal data, implementing more sensitive posterior predictive tests may prove fruitful. Future analyses could use matching distance instead of RF distance, since the former has more reliable statistical properties. Another approach could combine information from multiple test statistics

using multiple regression or principle components analysis and use this information to exclude genes with more model violation. This filtering approach may be extremely useful, since many datasets are already too large for analysis in popular software packages and reducing noise and bias should improve the accuracy of results.

The field of phylogenetics has shifted from being data-limited to being method- and computation-limited. Developing more advanced statistical approaches, making use of existing information, and improving software efficiency are necessary to obtain reliable results in a reasonable amount of time. Additional work is needed to enhance phylogeneticists ability to reduce bias and detect systematic error, but approaches such as informed priors and posterior predictive simulation may offer a promising way forward.

Bibliography

Andersen J.J., Nelson B.J., Brown J.M. 2014. EmpPrior: using empirical data to inform branch-length priors for Bayesian phylogenetics. Submitted.

Bogdanowicz D., Giaro K. 2012. Matching split distance for unrooted binary phylogenetic trees. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 9:150-160.

Bolker, B. 2010. bbmle: Tools for general maximum likelihood estimation.

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution, 19: 11711180.

Brown J.M., ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25: 537-538.

Brown J.M., Hedtke S.M., Lemmon A.R., Lemmon E.M. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. 59:145-161.

Brown, J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Systematic Biology syu002.

D'Erchia A.M., Gissi C., Pesole G., Saccone C., Arnason U. 1996. The guinea-pig is not a rodent. Nature 381:597600.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:19691973.

Efron, B. 2013. Bayes theorem in the 21st century. Science. 340:1177-1178.

Ekman S., Blaalid R. 2011. The devil in the details: interactions between the branch-length prior and likelihood model affect node support and branch lengths in the phylogeny of the Psoraceae. Syst. Biol., 60:541-561.

Gamble T., Berendzen P.B., Shaffer B., Starkey D.E., Simons A.M. 2008. Species limits and phylogeography of North American cricket frogs (Acris: Hylidae). Mol. Phylogenet. Evol.. 48:112-125.

Gatesy J., O'Leary M.A. 2001. Deciphering whale origins with molecules and fossils. Trends in Ecology & Evolution 16:562570.

Gauthier J.A., Kearney M., Maisano J.A., Rieppel O., Behlke A.D. 2012. Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. Bulletin of the Peabody Museum of Natural History 53: 3308.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2004. Bayesian Data Analysis, 2nd Ed. Chapman & Hall/CRC, New York.

Goldstein M. 2006. Subjective Bayesian analysis: principles and practice. Bayesian analysis. 1:403-420.

Hedke S.M., Stanger-Hall K., Baker R.J., Hillis D.M. 2008. All-male asexuality: origin and maintenance of androgenesis in the Asian clam Corbicula. Evolution. 62:1119-1136.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27:570580.

Hillis D.M. 1987. Molecular versus morphological approaches to systematics. Annual Review of Ecology and Systematics 23-42.

Hunter J.D. 2007. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9:9095.

Hurvich C.M., Tsai C.L. 1989. Regression and time series model selection in small samples. Biometrika. 76:297-307.

Jukes T., Cantor C. 1969. Evolution of protein molecules. Pages 21-132 in Mammalian Protein Metabolism. (H. Munro, ed.). Academic Press, New York.

Kass R.E., Wasserman L. 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association. 91:1343-1370.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16:111-120.

Lakner C., van der Mark P., Huelsenbeck J.P., Larget B., Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol., 57:86-103.

Larget B.R., Kotha S.K., Dewey C.N., An C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910-2911.

Leaché A.D., Mulcahy D.G. 2007. Phylogeny, divergence times and species limits of spiny lizards (Sceloporus magister species group) in western North American deserts and Baja California. Mol. Ecol. 16:5216-5233.

Liang L.J., Weiss R.E., Redelings B., Suchard M.A. 2009. Improving phylogenetic analyses by incorporating additional information from genetic sequence databases. Bioinformatics, 25:2530-2536.

Lin Y., Rajan V., Moret B.M. 2012. A metric for phylogenetic trees based on matching. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 9:10141022.

Losos, J.B., Hillis D.M., Greene H.W. 2012. Who speaks with a forked tongue? Science 338:14281429.

Marshall D.C. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol., 59:108-117.

Naylor G.J., Brown W.M. 1997. Structural biology and phylogenetic estimation. Nature 388:527528.

Naylor G.J., Adams D.C. 2001. Are the fossil data really at odds with the molecular data? Morphological evidence for Cetartiodactyla phylogeny reexamined. Systematic Biology 50: 444453.

Nelder J.A., Mead R. 1965. A simplex method for function minimization. The Computer Journal, 7:308-313.

Nowak M.D., Smith A.B., Simpson C., Zwickl D.J. 2013. A simple method for estimating informative node age priors for the fossil calibration of molecular divergence time analyses. PloS ONE, 8:e66245.

Paradis E., Claude J., Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289-290.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 5:e9490. doi:10.1371/journal.pone.0009490.

R Core Team. 2013. R: A language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing. Available from: `http://www.R-project.org/`.

Rambaut A., Drummond A.J. 2009. Tracer v1.5.0 [Internet]. Available from: `http://beast.bio.ed.ac.uk/Tracer`.

Rannala B., Zhu T., Yang Z. 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29:325-335.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131147.

Ronquist F., Huelsenbeck J.P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol., 61:539-542

Schliep K.P. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics. 27:592-593.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:15691571.

Sullivan J., Swofford D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. Journal of Molecular Evolution 4:7786.

Swofford D.L. 1991. When are phylogeny estimates from molecular and morphological data incongruent. Phylogenetic analysis of DNA sequences 295333.

Symula R., Keogh J.S., Cannatella D.C. 2008. Ancient phylogeographic divergence in southeastern Australia among populations of the widespread common froglet, Crinia signifera. Mol. Phylogenet. Evol. 47:569-580.

Tavare, S. 1986. Some probabilistic and statisical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57-86. 17:368-376.

Waddell P.J., Ota R., Penny D. 2009. Measuring fit of sequence data to phylogenetic model: gain of power using marginal tests. J. Mol. Evol. 69:289299.

Wendel J.F., Doyle J.J. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. Molecular systematics of plants II 265-296. Springer US.

Wetterstrand K.A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed June 10 2014.

Whelan S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol. Biol. Evol. 25:1683-1694.

Wickham H. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

Wiens J.J., Hutter C.R., Mulcahy D.G., Noonan B.P., Townsend T.M., Sites J.W. Jr., Reeder T.W. 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biology Letters 8:10431046.

Yang Z., Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol., 54:455-470.

Zhang C., Rannala B., Yang Z. 2012. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. Syst. Biol. 61:779-784.

Zwickl D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

Vita

Bradley J. Nelson completed his undergraduate studies in mathematics and biology at Seattle Pacific University in August 2010. In June 2011, he came to Louisiana State University to pursue graduate studies in biology. He is currently pursuing a Master of Science degree in biology, which will be awarded in August 2014.